Prediction and regression

Statistics for Social Scientists I, L7

Bur, GJM

2024-07-01

1 Prediction generally

At the start of the previous lecture, we discussed the difference between working out an association between variables and *predicting* the value of one variable based on another. They are of course closely related; I'll discuss the exact connection later. For now, let's move to the basics of prediction.

Here, predicting does not mean necessarily guessing about a *future* event, although it can. It refers more *generally* to the process of learning about the unknown, information *reduction*, or discovering causal relationships. For example, if I work in marketing and want to figure out what a customer's credit score is likely to be, information that might be expensive to track down or which the customer may not know, it might be possible to guess their score based on information that they are willing to provide—age, monthly income, education, etc. (this is basically a learning/information compression example).

What about a more sociological example? Well, *under certain circumstances* (complicated ones which we can't get too deep into), we can say that if X predicts Y once we adjust for other variables (called *covariates*), X probably causes Y. We can also make steps towards this process of causal discovery with correlation, but regression is a more natural choice when we have a variable we think of as the outcome. Regression also transmits information about association, as well, so it does most of what correlation does; we'll see later the very close connection between them.

2 Optimal prediction is surprisingly easy: conditional means

Remember before when I mentioned that if we have a very simple statistical model of the form $Y = k + \epsilon$, where k is some constant and ϵ is a random noise term, that the best[1] possible guess, the best value for k, is the mean? It turns out that this generalizes very naturally. Let's take the mean of that loss function, the sum of squared "errors" (the gap

[1] Well, given the most common measure of "best", which means minimizing a particular kind of loss function, $L(\theta) = \sum_{j=1}(y_j - \theta_y)^2$. Also, note that the use of Greek ϵ violates our rule that we mentioned above about Greek letters generally being parameters, constants that characterize a distribution; here it is a random variable, but this notation is very standard, so I'll keep it.

between an observation and our guess), and give it its standard name, the *mean squared error*, $MSE: \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$. The "hat" on y (technically a *caret* for you grammar fans) means "prediction". Here it just means a generic prediction, of which $k = \bar{y}$ or $k = \text{median}(y)$ are examples. Those predictions, however, are just single numbers that don't take any other values x_j into account.

Now, let's think about allowing for a more general statistical model that might make use of information from variables like X and think about how we might minimize the *MSE*. Perhaps we think that someone's height or income can be predicted by something more than its own average. Maybe our model at the population level looks generally like this: $Y = \beta_0 + f(X) + \epsilon$. All this amounts to is a very general statement of the following type: "weight is some function of height, plus a constant and random noise".

Notice that we're making a lot of assumptions when we write our model that way. First, that just one predictor is necessary, which won't always be so plausible, although in today's example, it's not so bad. We'll see at the very end of the course that we can allow more variables in, if we want. Second, and relatedly, we assume that the error term is truly random, meaning that it isn't related to the value of X; our predictions aren't better or worse anywhere in the input space (here, that means "along the x-axis"). A third, though minor, assumption is that ϵ has mean zero, although this is kind of trivial since if our model turned out to have errors, we could just take their average and put them into the intercept (meaning that it is possible prediction can be systematically wrong but have no overall error).

I'm relegating the proof to the appendix, but intuitively, it seems clear that if the mean minimizes the loss overall, the *conditional* means—i.e., the mean taken at each possible value of our predictor—should give the best possible prediction given our predictor variables. This is, in fact, true. If we include the *correct mix of variables*, we actually can't do better than using the conditional means, *period.*

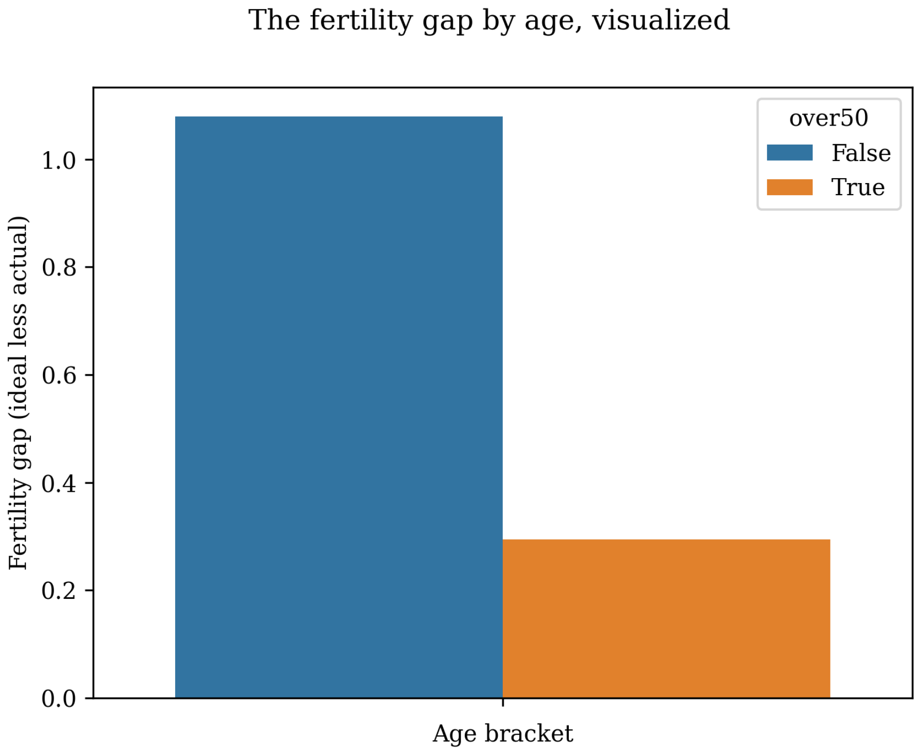
Even better, we've actually already seen these conditional means before, throughout the notes. For example, in L3, when we calculated the mean fertility gap based on the age of individuals, we were finding conditional means. For the sake of writing down concise notation, let's just call these conditional means $\bar{y}_h$, where h just indexes the value of the predictor.

For example, we found $\bar{y}_{\text{over 50}}$ and $\bar{y}_{\text{50 or younger}}$. As a reminder, here is what that table looked like after we made our fertility gap variable, and here are three ways to see the difference visually. Note that I also include the conditional five number summary and variance.

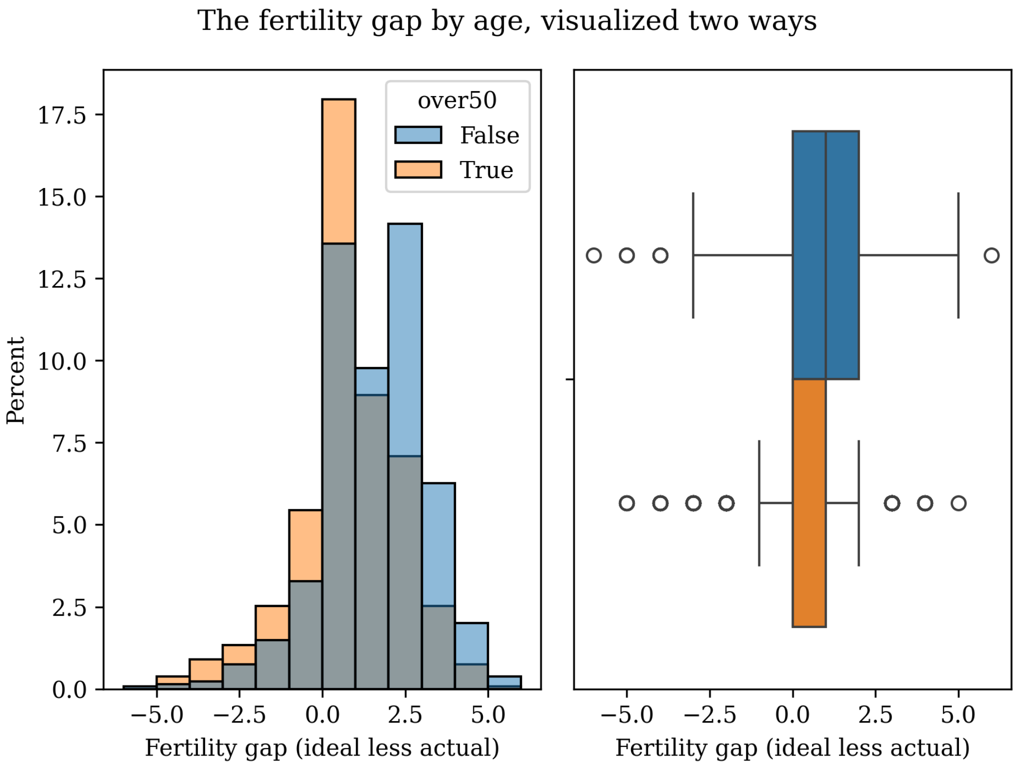
over50	count	mean	std	min	25%	50%	75%	max
False	699	1.08011	1.58228	-6	0	1	2	6
True	643	0.293935	1.5902	-5	0	0	1	5

Here's what that looked like graphically in the simplest format, a conditional bar graph. The Stata code to get something like this, using some easily-accessible data (`sysuse auto`) is...

```
graph bar price, over(foreign) title("The fertility gap by age, visualized") ///
    ytitle("Fertility gap (ideal less actual)") b1title("age bracket")
```



Here were two alternatives that don't directly show the mean but get at the idea of a conditional distribution. You can see my project do-file if you're interested in making a more com-



plex graph like this.

The big takeaway is this: if you have a qualitative predictor and a quantitative outcome *or* a discrete-quantitative predictor with few values, you should just find conditional means—and you should also generally report the five-number summary as well.

The goal of the rest of this lecture will be to help us fill in chart below for “quant. outcome, quant. predictor”.

	quant. outcome	**qual. outcome**
qual. predictor	conditional means	?
quant. predictor	?	?

3 Conditional means: what to do with quantitative predictors?

3.1 The problem

Now, suppose that we have a quantitative predictor. We *could* try to take conditional means of it. Perhaps we want to predict a person’s education from their father’s using GSS data. How well will that work out? Let’s try a table of conditional means alone.

paeduc	mean	count
0	9.5	42
1	12	1
2	11.8333	12
3	11.8485	33
4	11.2	15
5	12.88	25
6	12.5625	64
7	13.0303	33
8	13.7232	112
9	13.4043	47
10	13.1852	54
11	13.82	50
12	13.9799	596
13	14.8	55
14	15.0368	136
15	14.9655	29
16	15.1597	238
17	16.3684	19
18	16.2769	65
19	15.6923	13
20	16.2917	48

Notice how very small some of our conditional n_h are. Do we really want to rely on tiny sample sizes in some cases? Also, note how there appears to be a pretty clear pattern in the data. Perhaps conditional means would be *inefficient* as well, in that they require us to provide a great deal of information just to summarize what might be a relationship of the form $y = mx+b$ (plus random noise). By the way, the bar graph would look correspondingly horrible, so I omit it.

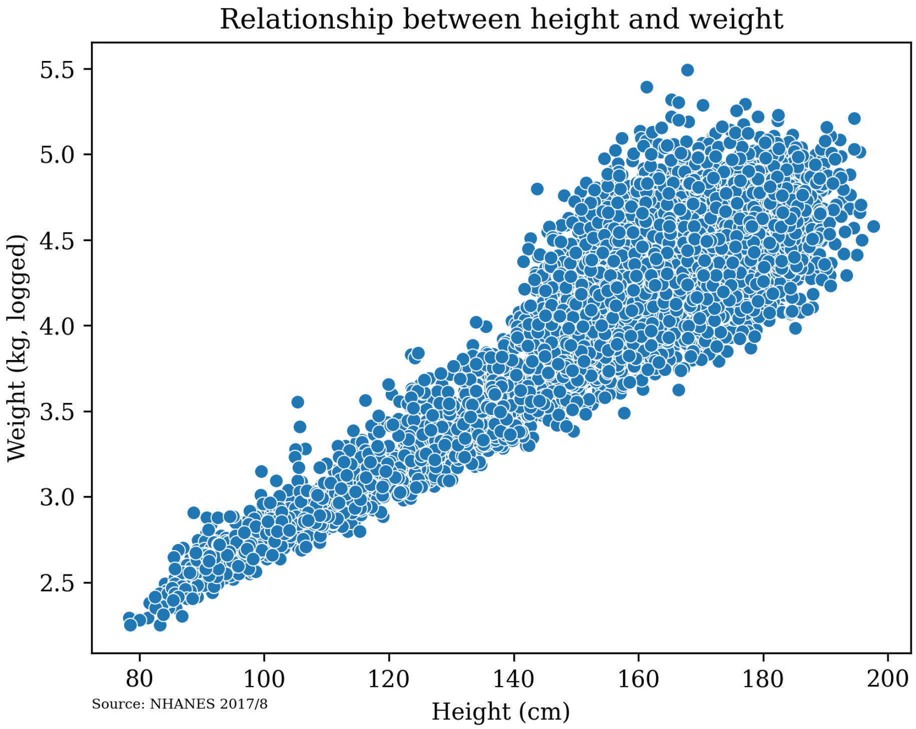
3.2 One solution: model conditional means as a line

Let's recapitulate out our problem here.

First, we might have a very small sample size in some regions of the "input space" (possible values of the predictor), but this may not reflect an absence in the population.

Second, our conditional means might be *inefficient* in the sense that they are by definition a very general kind of **regression function**, written $\mathbb{E}[Y|\mathbf{X}]$ at the population and meaning "the conditional mean of Y given $\mathbf{X}$ which might be comprised of multiple predictors"...but this function might actually be something very easy to describe with a line, like $Y = \beta_0 + \beta_1 X_1 + ... \beta_k X_k + \epsilon$. Here we'll just deal with one predictor, one X, at a time.

This second assumption is a big assumption, to be clear! However, in some cases it is pretty clearly an OK. Let's return now to our NHANES example of height and logged weight. Clearly we can characterize this situation pretty neatly with a linear equation. *But how can we write out the best line*? Even if eyeballing it is not too hard in this situation, the problem gets more intractable in three dimensions or more.



4 Regression: a geometric derivation

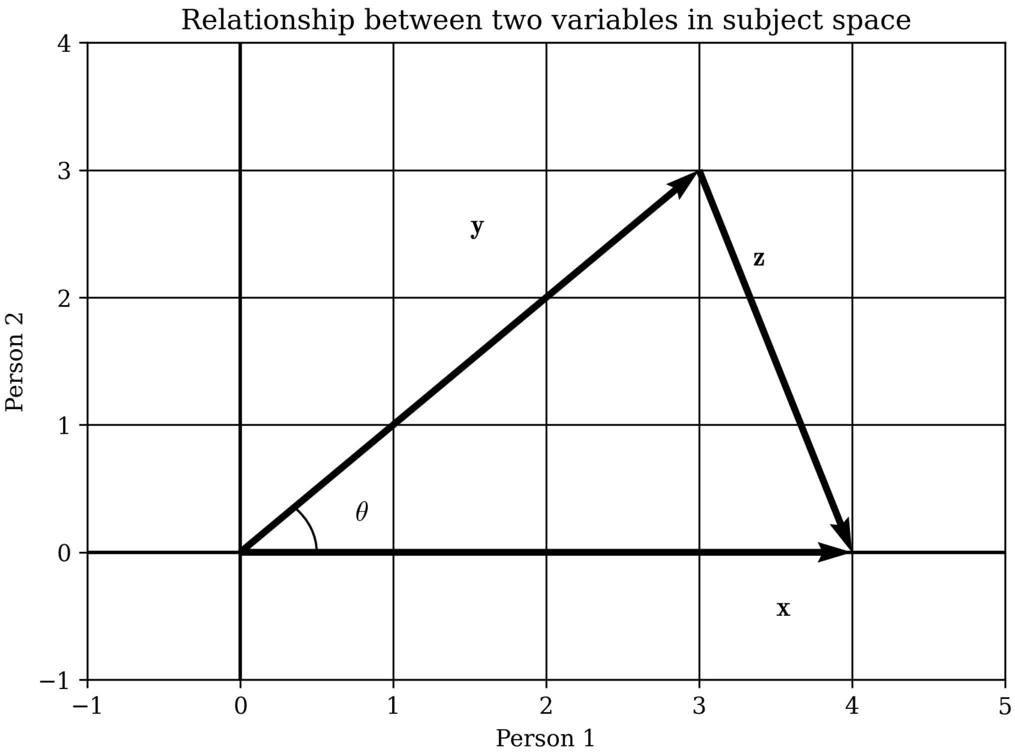
So, here is our task: find the best possible line of the form $Y = \beta_0 + \beta_1 X$ in our sample to describe the data. *We'll generally write the sample estimate of this line as $y = b_0 + b_1 x$.*[2]

Let's again use an alternate geometry to motivate this problem, making it incredibly easy to solve (I think). Here's what I said last lecture about this *subject space*, which we use to represent our sample data. It bears repeating.

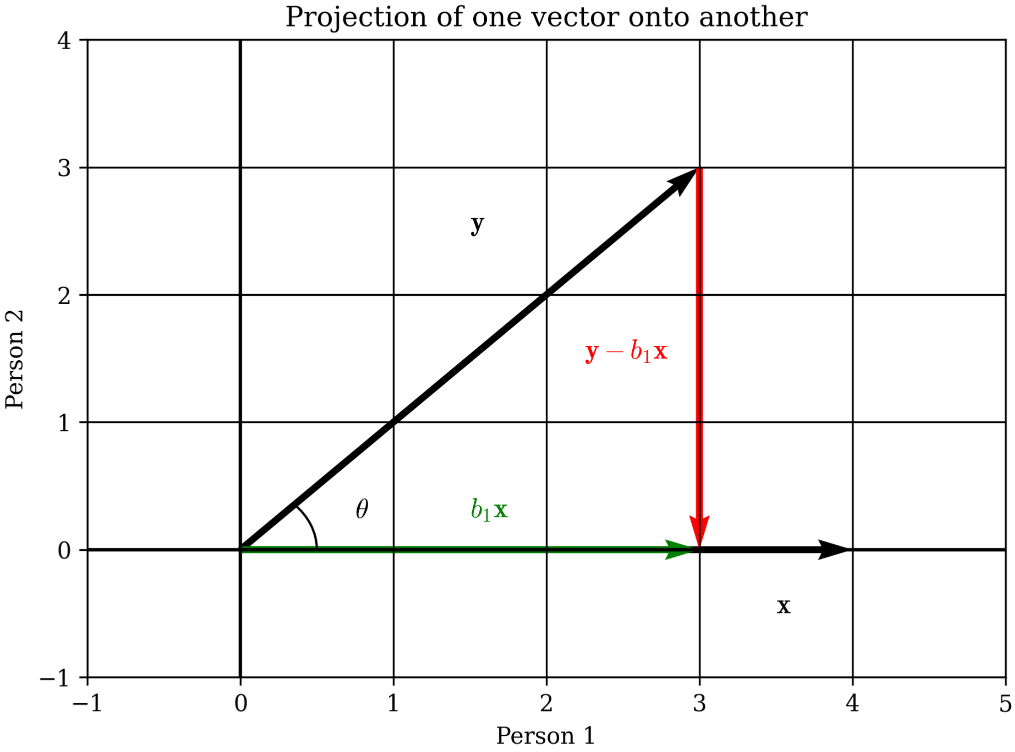
> Recall that above, we were working in what is called *variable space*: the axes of that two dimensional space each represent a variable. But, recall that we also can picture our variables in what is called *subject space*, where each axis is a person. Recall that this is an especially natural tool if we think about *centering* the original scores; so, for example, the whole vector of scores $\mathbf{y}_c = \mathbf{y} - \bar{\mathbf{y}}$.
>
> Here is a picture of how that looks in just two dimensions. On the variable x person 1 had a (centered) score of 4; person 2 had a score of zero (centered; I won't write that explicitly anymore). So, that variable is the vector $(4, 0)$ (a

[2] You might also see the *betas* with hats on them, like so: $\hat{\beta}$. This is more technically correct, but it is kind of hard to look at, and I've already used the tradition of using Roman letters for sample estimates, like s for σ and r for ρ.

vector is just a point drawn as an arrow). On variable y, person 1 had a 3, while person 2 had a 3 also, so that is the vector $(3, 3)$.



The idea here is very simple. **I want to find some multiple of x in the sample to get it as close to as possible to y (stretching a vector representing x in subject space by *a constant* b_1*, the sample estimate of* β_1, to produce a list of predictions $\hat{\mathbf{y}}$ is equivalent to drawing a line $\hat{y} = b_1 x$ in variable space).**



4.1 Regression slope: basic formula

Now, the generalized law of cosines gives us our answer for how to find the perfect sample estimate b_1 of our population regression slope β_1. We see above that our error vector is smallest if it is perpendicular to our predictor. Let's assume for now that our intercept β_0 is zero; we'll motivate this in a moment.

The law of cosines tells us that for two vectors to be perpendicular (or *orthogonal*, a term more common in this context), their "dot product" must be zero.

What is this scary-sounding "dot product"? You'll be very pleased to learn that it is something we have already calculated: for two vectors $\mathbf{w}$ and $\mathbf{z}$, it is $\mathbf{w} \cdot \mathbf{z} = \sum_{j=1}^{n} w_j z_j$.

This looks very much like a correlation's numerator, especially if we consider that these vectors in the picture represent *centered* versions of our variables: then, for our centered error and outcome vectors to be orthogonal, we have $[(\mathbf{y} - \bar{\mathbf{y}}) - b_1(\mathbf{x} - \bar{\mathbf{x}})] \cdot (\mathbf{y} - \bar{\mathbf{y}}) = \mathbf{0}$, which can be written in summation notation as $\sum_{j=1}^{n}[(y_j - \bar{y}) - b_1(x_j - \bar{x})](y_j - \bar{y}) = 0$.

So, let's make our residual vector or in-sample error as small as possible by making it orthogonal to our predictor (for the simple reason that the shortest distance between two points is a straight line; our error vector is smallest if it is a straight, not oblique, line from the scaled predictor to the outcome).

First, write both quantities in centered form; for the in-sample error, it is $[(y_j - \bar{y}) - b_1(x_j - \bar{x})]$.

$$\begin{aligned} \sum_{j=1}^{n}[(y_j - \bar{y}) - b_1(x_j - \bar{x})](x_j - \bar{x}) = 0 \\ \sum_{j=1}^{n}(y_j - \bar{x})(x_j - \bar{x}) = b_1 \sum_{j=1}^{n}(x_j - \bar{x})^2 \\ \frac{\sum_{j=1}^{n}(y_j - \bar{y})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = b_1 \end{aligned}$$

4.2 Regression slope: close relation to correlation

Even better news! Doesn't this look *very* similar to our correlation coefficient?

$$r = \frac{\sum_{j=1}^{n}(y_j - \bar{y})(x_j - \bar{x})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}}$$

In fact, we can just multiply the correlation coefficient by the ratio $\frac{s_y}{s_x}$ to get b_1. The algebra is easy but tedious. *I encourage you to try this on your own; see appendix for proof.*

4.3 Regression slope: centered example, motivating the intercept

Let's tarry no further. What should the *slope of our line of best fit* for the height/logged weight data above? Recall that we had a correlation of $r = 0.8942$ and the following summary statistics (I'm being a bit more precise than normal here since one of our variables is logged)...

	ht	logwt
mean	156.5864	4.1066
std	22.2551	0.5601

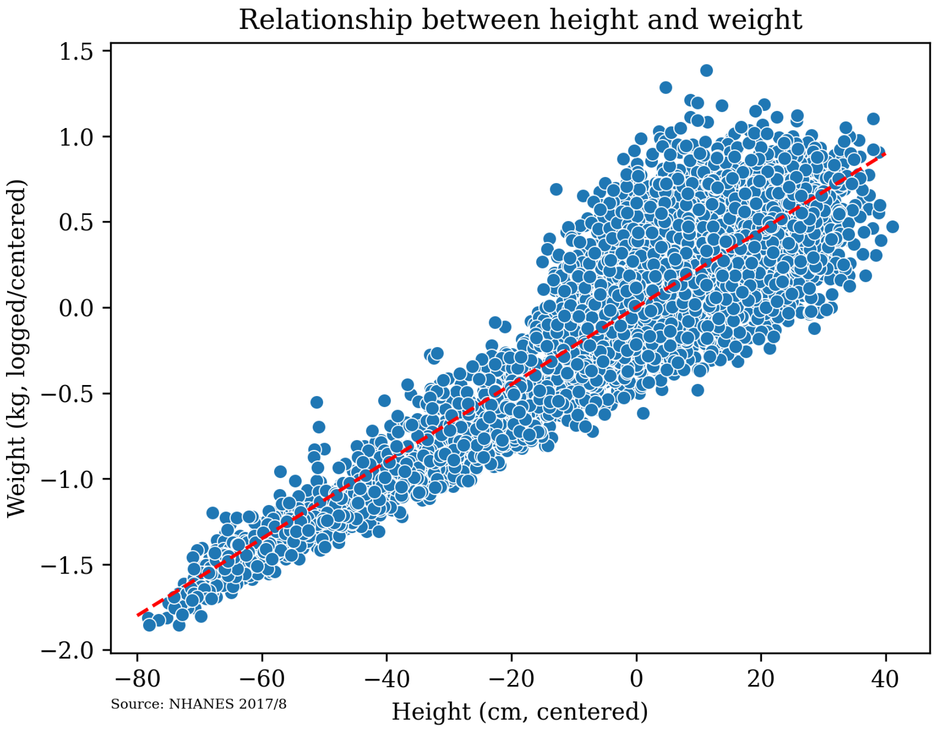
So, what is the slope of our line of best fit? Simply, $b_1 = r \cdot \frac{s_y}{s_x} = 0.8942 \cdot \frac{0.5601}{22.2551} = 0.0225$.

Now, what should our estimate of the **intercept** (b_0 in the sample; β_0 in the population) be? I can give you a formula here, but what if I motivate it a bit?

Recall that above we had centered data. The *centering* operation is very often useful, as we've seen. It changes little except very basic summary statistics. Then, we would take our population model, $Y = \beta_0 + \beta_1 X_1 + \epsilon$, and rewrite it. To find the model for centered data, we can write $(Y - \mu_Y) = \beta_1(X_1 - \mu_X) + \epsilon$ by simply taking the deviation of both sides from their mean.[3] The intercept drops out because its mean is just itself. We can rewrite our

[3] By linearity of expectations, the mean of the right-hand side is just the sum of the means of each variable: ϵ has mean-zero by construction, and β_0 is constant, so it is its own mean and thus subtracts itself out.

sample regression equation accordingly as simply $(y - \bar{y}) = b_1(x - \bar{x})$. Then, we have no need for a constant term. When our centered predictor $(x_j - \bar{x}) = 0$, so must y by construction. Our regression line for centered variables looks like this.



Source: NHANES 2017/8

So, what if we *don't* have mean-zero variables? Well, our slope remains correct—geometrically, you can think about the process of restoring the means in variable space as just shifting all of our points right and up (assuming both had positive means); this kind of translation can't affect the slope, of course.

Rearranging our population model so that we now state it in terms of *uncentered* variables, we have the following: $Y = \mu_Y + \beta_1 X - \beta_1 \mu_X + \epsilon$. Our sample regression line, accordingly, is $y = \bar{y} + b_1 x - b_1 \bar{x}$.

This tells us that, with uncentered variables, our best prediction is the slope from the centered model multiplied by x, but with an extra $\bar{y} - b_1 \bar{x}$ added on. We call it our **intercept and denote it** b_0 in the sample.

So, generally…

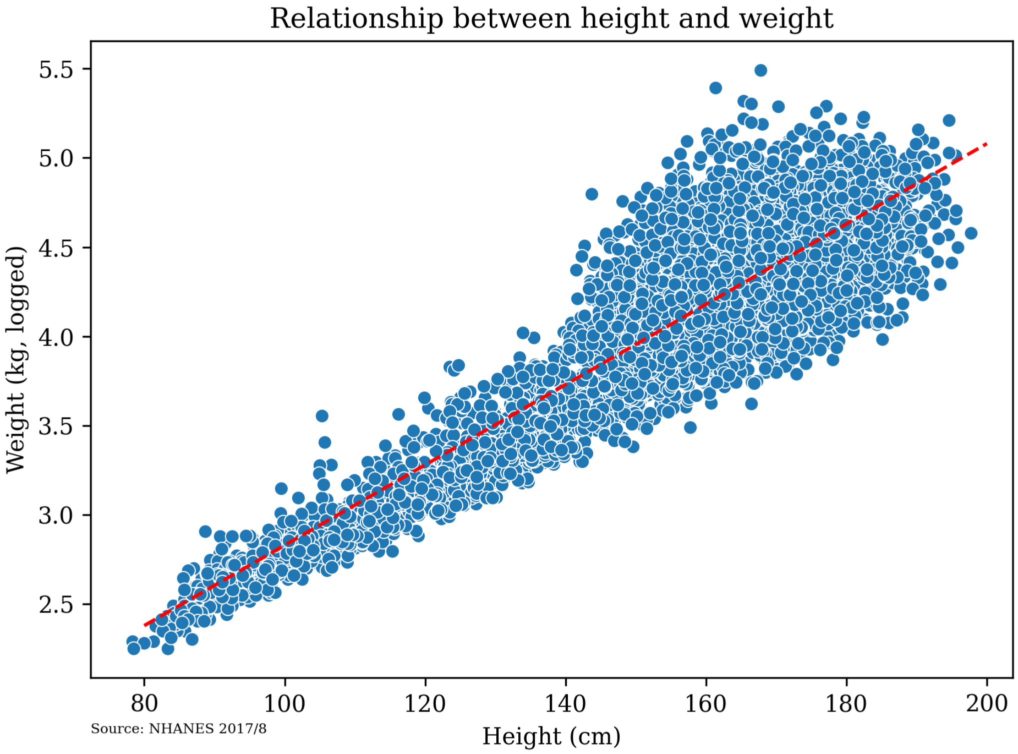
$$b_0 = \bar{y} - b_1 \bar{x}$$

Our intercept for this model is therefore $4.1066 - 0.0225 \cdot 156.5864 = 0.58$. So, our regression line now looks like this: $y = 0.58 + 0.0225x$.

4.4 Regression line interpretation: example

We interpret this as saying that for a unit increase in one's height, their (logged) weight increases by about 0.0225 kilograms.

Here, as is often the case, our intercept isn't meaningful because a zero weight is not meaningful. *You'll see this situation a lot when you're first starting out and need to write linear models for situations that aren't strictly linear.*[4]



Source: NHANES 2017/8

5 Regression diagnostics

Now that we've worked an example—the actual computations here are very simple if we have the correlation coefficient, standard deviations, and means handy, so I'm punting on more computational practice until the end—let's talk about how to assess our model.

Recall the two main assumptions of our model.

Firt, we assumed that we we included all relevant variables in the model. Failure to include the correct variables in the model makes our slope wrong in terms of the causal effects of X, but it can still be valid descriptively. We'll see an example later.

Second, we assumed that we had the right functional form. Since we're working here with

[4] Here, we have a nice "out" in that we can exponentiate both sides, as we'll see later. For example, here, we have $y = e^{0.58+0.0225x}$.

linear model, this means that the actual data-generating process actually work like our model says it does: $Y = \beta_0 + \beta_1 X + \epsilon$. The predictor X itself can be something like a square or logarithm of some underlying variable (indeed, that's what we've already seen), but the model must be linear in its *parameters*: if we have multiple variables, the model is linear if and only if we just multiply the variables by constants and then add to get predictions. *Other types of functional forms are possible, and our linear model is misspecified if reality is not linear like this.*

In what follows, I first address the question of quantifying out model fit—how good are our predictions?—which can be seen as a means of sussing out whether we have the right variables in our model and the right functional form. Then, I discuss head-on some more-direct, intuitive ways of looking at whether our assumptions are correct and what to do if they aren't.

5.1 Predictions, errors, the mean squared error, and model fit

Let's talk about model fit. This stuff is actually very interesting, rather than just being a bunch of nitpicking; it's quite geometrically beautiful.

First: how good of a fit is our model? Well, unsurprisingly, we'll again use the mean squared error as a measure. We saw before that conditional means minimized the MSE, and if our model is truly a linear one, our simple formula $b_1 = r \cdot \frac{s_y}{s_x}$ is the regression slope that minimizes in-sample errors.[5]

These in-sample errors, training errors, or residuals are the gap *in the sample* between our prediction, written $\hat{y}_j = b_1 x_j + b_0$, and our actual values, written of course y_j. Note that with linear regression we almost always have points with the same input value x_j so that many different individuals share the same prediction.

It is important to distinguish in-sample errors or residuals (same thing) from the out-of-sample errors, the latter of which combine error arising from the fact that our sample model is not only never-quite-perfect for data we *do* have but also is itself only a guess at the true model parameters (because estimated from sample data).[6] **We will use the "hat" notation $\hat{y}_j$ for our *prediction* for person j.** This is just equal to the model's value at person j's value. Then, **the error for person j is simply** $err_j = y_j - \hat{y}_j$. This is how much we miss at that point.

Now, a global measure of error would simply be the mean squared error, that same loss function that we've used time and again, sometimes as a total, sometimes as a mean, sometimes with the root taken—but always that same sum of squared deviations appears. Pretty handy how this stuff doesn't change that much, huh?

In a regression context, the (near) average of the square deviations is called the "mean squared error" and standard statistical software reports it or its root, the "root mean squared

[5] Making the error vector perpendicular to the predictor also minimizes it; you can formally write this as a minimization problem and solve the regression problem with calculus—but I think the linear algebra approach is even prettier).

[6] Another way to say this is: our sample model is the right model for the "population" that is our sample, assuming our theoretical selection of variables and their functional form is correct. It still won't be perfect, of course. But, for data points not in our sample, we also have to deal with the fact that the model is not the right model for the actual population, just a guess at it. So it will miss points not in our sample for that reason as well.

error". These are basically the variance and standard deviation of the data but with the prediction for a person swapped in where we had the mean before (and recall that the mean is our most basic predictor, so it's not even that different). *The sum of squared errors is also the length of that error vector we saw before!*

Just make sure that you remember that this is the *in-sample* error, which is why it is sometimes called a "residual" (to distinguish it clearly from out-of-sample error). *Older books and software are frustratingly inconsistent in this language, so I will be ultra-precise: residual = in-sample error =/= out-of-sample error.* In general, to interpret "error", you need to know the context.

$$
\begin{aligned}
SS_E &= \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \\
MSE &= \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \\
RMSE &= \sqrt{MSE}
\end{aligned}
$$

5.2 The "coefficient of determination", aka R-squared

The MSE is a nice quantity to have around, but it, like all variance-forms, is not a quantity we can just compare blindly across contexts. Is there a form of model fit that is perhaps more like a percentage, which can at least sort of be compared? Yes. We can scale the MSE by dividing it by the overall variation, $\frac{MSE}{s_y^2}$; this quantity is then conventionally subtracted from 1 to give it a positive interpretation: **how much of the total variance in the outcome does our model explain?** Logically, the two must sum to 1, another of our probability/counting rules. This also has a very pleasing geometric interpretation.

First, let's just write out this ratio, whose name is R-squared or R^2. You might wonder if this is related to the correlation coefficient. It is! (Why it is not just written r^2, I don't know). In what follows, remember that the numerator of the variance, $\sum_{j=1}^{n} (y_j - \bar{y})^2$, is often referred to as the *total sum of squares*, or SS_T. *Recall that this is just the squared length of a vector if we draw a centered version of the data as a vector.*

$$
\begin{aligned}
R^2 &= 1 - \frac{MSE}{s_y^2} \\
&= 1 - \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} \\
&= 1 - \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{SS_T} \\
&= 1 - \frac{SS_E}{SS_T}
\end{aligned}
$$

The final line has a very nice interpretation in the centered subject-space picture we've been using off and on throughout all the notes: it is just one minues the ratio of the squared length of the vector representing our in-sample errors to the squared length of the vector representing our variable. Pretty easy to interpret, no? Statistical software reports these quantities when you carry out a regression.

Even better, consider the fact that since the in-sample error vector is totally orthogonal (perpendicular) to the vector representing our predictor.[7] This means that by the Pythagorean theorem, we can immediately calculate R^2 by simply finding the the squared length of that prediction vector, conventionally called the "model sum of squares" (SS_M), and dividing it by the total squared length of our outcome to get our model fit.

The squared length of our prediction vector is called the "model sum of squares" and written SS_M. The formula for it is as follows:

$$SS_M = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2$$

Then, **we can say that by the Pythagorean theorem, our quantity R^2 is just the ratio of our prediction vector's squared length to the total sum of squares! So, we can also write $R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_T - SS_E}{SS_T} = \frac{SS_M}{SS_T}$.**

By the way, for the exact way that these formulae for sums of squares connect to the geometry, see the appendix.

5.2.1 Correlation connection

So, do we have to calculate this beastly thing in order to find our model fit? Actually, no. Believe it or not, there is a major shortcut!

Notice that the ratio of the squared length of our predictors to the total squared length of $\mathbf{y}$ is just the ratio of the adjacent side of the angle between them, $\theta_{\mathbf{x},\mathbf{y}}$, to the hypotenuse of a right triangle. That means that this ratio is … the square of $\cos(\theta)$! If you recall from last lecture that the correlation coefficient is the cosine of θ, we can then just take the square of the correlation coefficient and obtain our model fit. **Hence the name "R-squared".**

5.3 Cross-validation

Another technique for evaluating our model fit, which goes back to the beginnings of statistics but which has unfortunately mostly been abandoned in social science (and rediscovered by machine learning), is that of **cross-validation**. The idea is simple and appealing, although computationally expensive; I quite honestly suspect that one reason that it was not fully pursued (in favor of complicated formulae) is that it seems "unserious" because it's so easy to understand.

Here is the idea. Partition your data into k different disjoint subsets (fancy math-speak for no overlaps, no repeats, everyone must go into one bucket). Fit your model on $k - 1$ of the

[7] And also the predictions, since they are just multiples of one another: $\hat{y}_j = b_1 x_j$ for all j.

sets, and then try to use it to predict the last set. Repeat until you've used every subset as the "holdout" or **test set** (the remaining $k-1$ sets are calling the **training set**). This gives us a guess at the out-of-sample MSE is then just the average of the squared residuals across this procedure.

5.4 Inspecting residuals

Remember that Normal curves describe the distribution of errors/random noise about a center quite well. So, our residuals or in-sample errors should be approximately Normal if they are truly *random*, meaning that our model explains everything systematic about our outcome (and so the "noise terms" are just that: unsystematic, uninformative noise). They might not be, though!

They should also generally be *conditionally* mean-zero, meaning that at each value of our predictor x, the mean of the residuals should be about zero. Recall that our method of estimating the regression function forces the *overall* in-sample error to be zero, *so this fact of zero overall residuals is never "proof" of anything good about our model.*

However, a substantial assumption that *can* be checked is the zero conditional mean assumption, which basically means that our functional form is correct and we've included all or most of teh relevant variables. The alternative is that we get overall zero-error in the sample by way of massive positive residuals in one region compensate for massive negative residuals in another reason (a guaranteed outcome if data are strongly non-linear).

Finally, they should also generally have the *same* conditional variance throughout the input space. The opposite of that situation is called "heteroskedasticity" ("different skew"), a horribly obscure term that you might hear smart people or wannabes drop.[8]

Good news: I won't subject you to the formal tests for these things. You should just know about them and be able to eyeball them.

5.5 Regression diagnostics: example

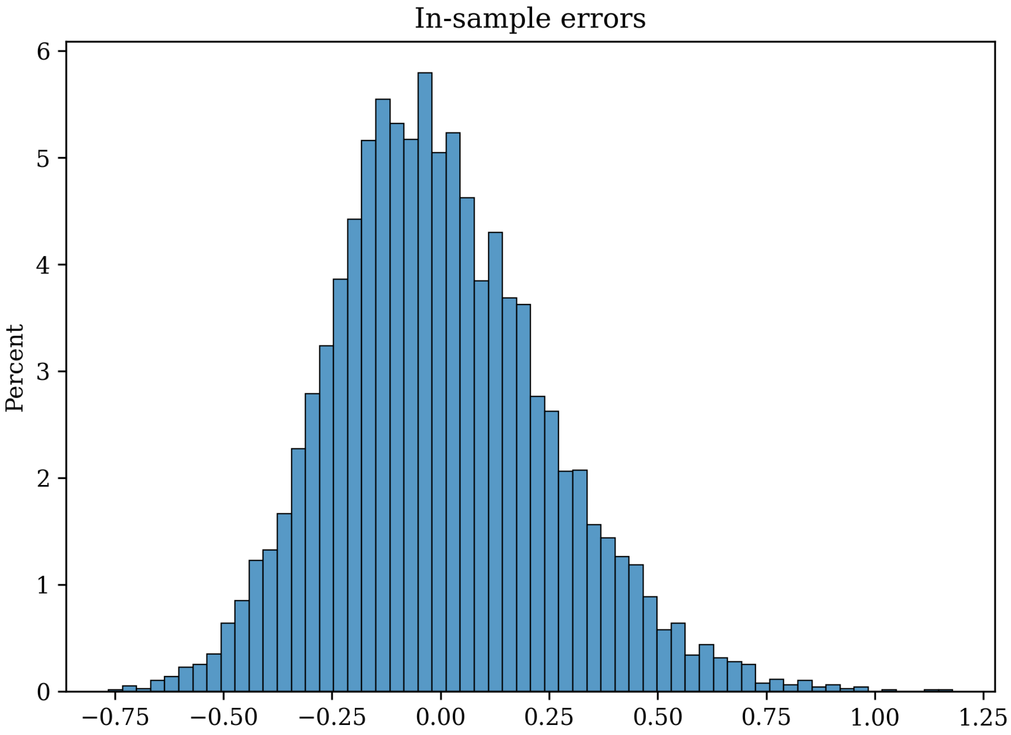
Let's conduct some diagnostics with our running NHANES example. We previously found that $y = 0.583 + 0.0225x$ was our regression line, using the following facts (now compiled into one convenient table). As a reminder, we found that $r = 0.8942$, and our table of means and standard deviations was...

	bmxht	logwt
mean	156.5864	4.1066
std	22.2551	0.5601

Let's immediately calculate R^2. It is simply $r^2 = 0.8942^2 \approx 0.80$. That means that about 80 percent of the variance in our outcome is explained by variance in our predictor. Recall that this is also the ratio of the model sum of squares to the total sum of squares, i.e. the squared length of our predictions as a percent of the squared length of our outcome (in subject space). So, we've done quite well here!

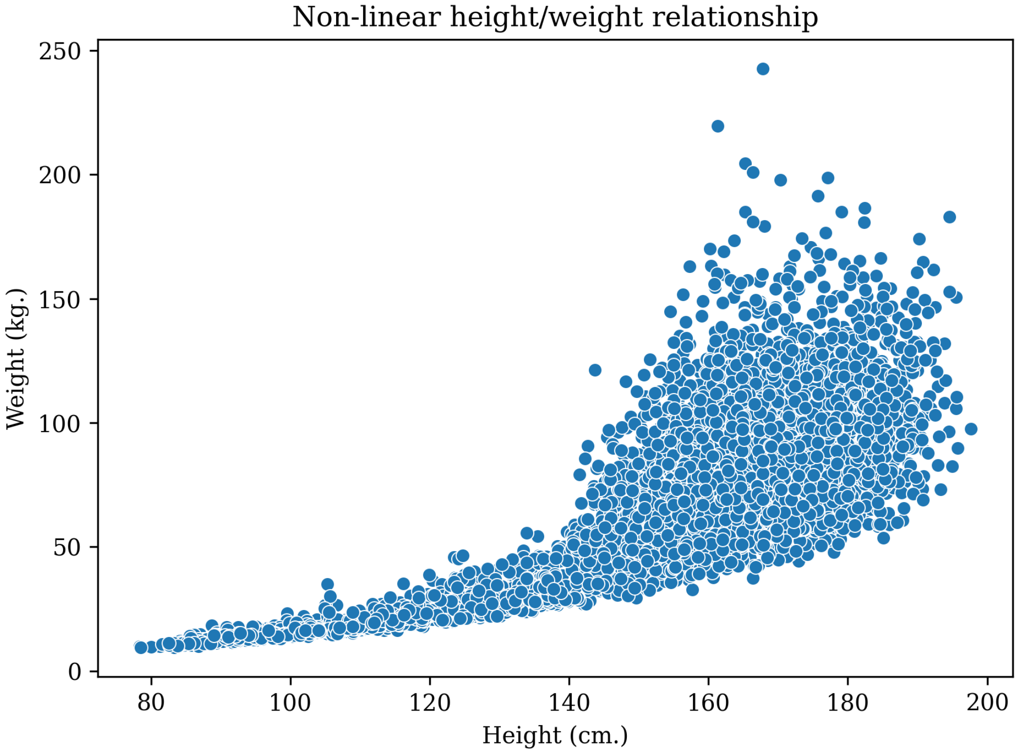
[8] My father, who is the chief financial officer at a software firm, once sent me a panicked text asking me to remind him what this meant before he had to reply to a "quant" at his job.

Now, let's examine the residuals themselves. Here is a histogram showing their distribution. These look pretty Normal: a good sign!



Source: Regression of logged weight on height using NHANES 2017/8

Now, let's do a little model comparison, which is where this stuff gets more interesting. Let's say that we want to try to examine a model with our *original* weight outcome, not the logarithm of weight. As a reminder, that looks like this.



Let's now fit a line to our *original, untransformed weight data* and compare it to the line we fit above *with a non-linear transformation of the outcome*. Here is a new table of summary statistics for weight, non-logged. The correlation between height and weight untransformed is $r = 0.7704$. Try to find the regression equation for the "untransformed" weight data.

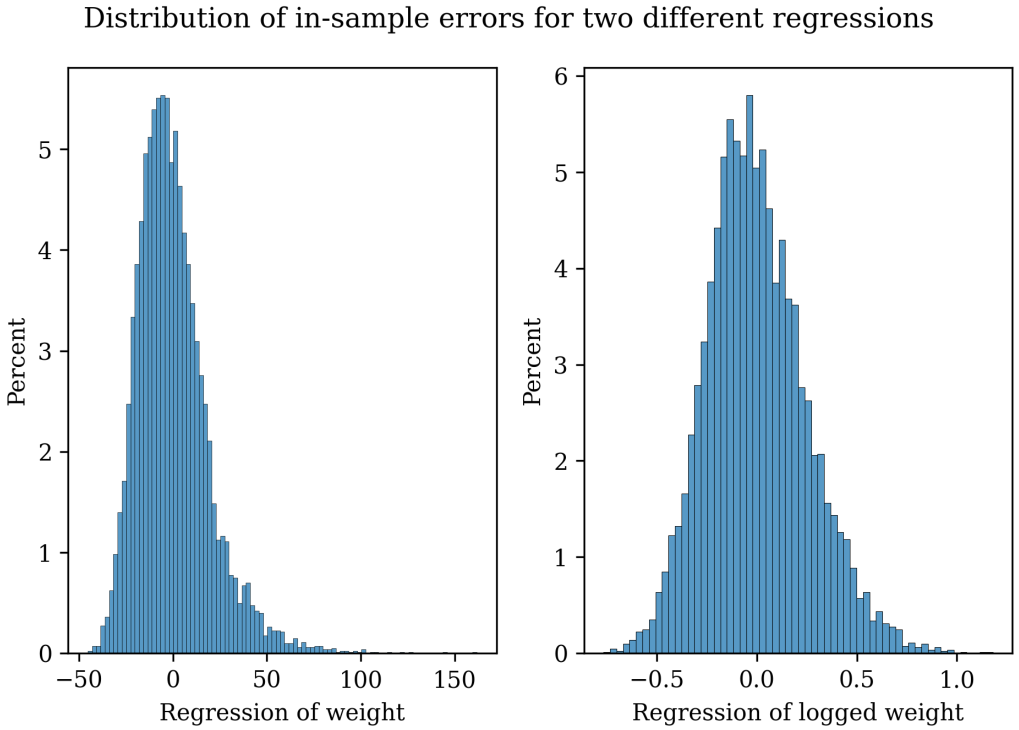
	bmxht	wt
mean	156.5864	69.0510
std	22.2551	30.3888

You should get the following.

$$\begin{aligned} b_1 &= 0.7704 \cdot \frac{30.3888}{22.2551} = 1.0520 \\ b_0 &= 69.0510 - 1.0520 \cdot 156.5864 = -95.68 \\ y &= 1.0520x - 95.68 \end{aligned}$$

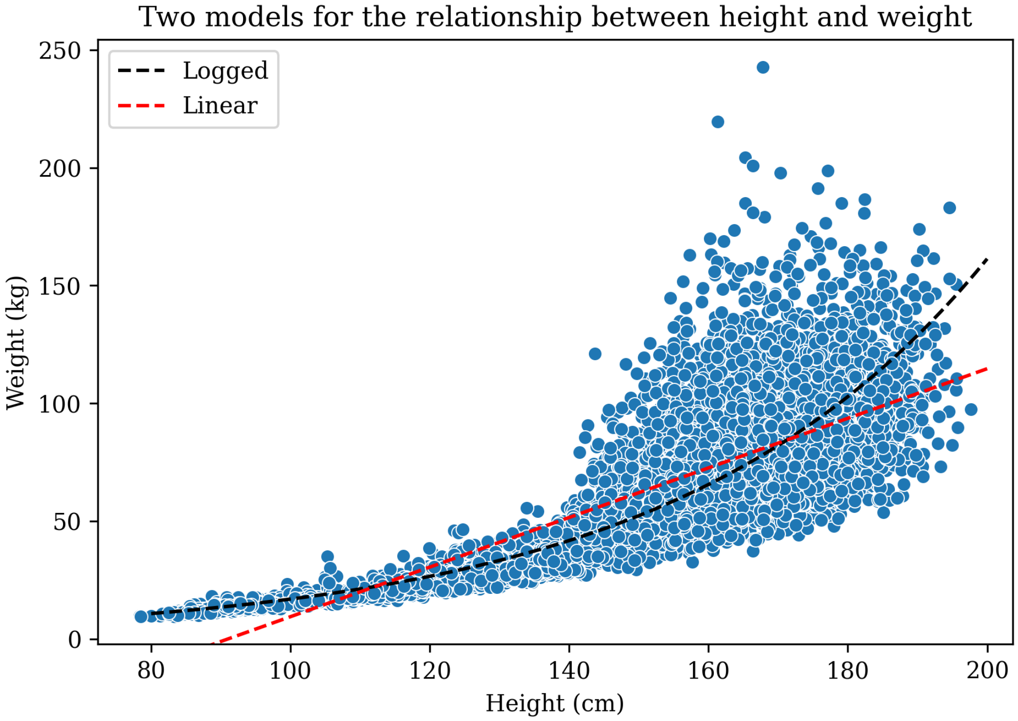
So, which model is better? Well, we could simply find the variance explained by each model. Our old model explained roughly 80 percent of the variation in the outcome; our new model explains only $0.7704^2 = 0.59$, or 59 percent, of the data. So, we should probably prefer the model with logged weight as te outcome.

We can also compare the residuals. Which look more Normal?



Source: NHANES 2017/8

To visualize the difference, let's get both models up on the original data. Here, I exponentiated both sides of the original model with the logged outcome so that they can be put on the same data. In which figure do we see greater spread about the regression line (heteroskedasticity)? In which figure do we see notable regions of the input space where the in-sample errors clearly have a non-zero mean in that spot?



Source: NHANES 2017/8

Finally, let's try our *k*-fold cross-validation. Below, I show the results. As predicted, our logarithmic model worked better.

test_set	log_slope	lin_slope	err_log_sq	err_lin_sq
3	0.90	0.77	345.21	414.94
1	0.89	0.77	389.19	433.73
4	0.89	0.76	401.96	385.36
5	0.90	0.77	376.81	430.33
2	0.90	0.77	365.54	409.70
mean	0.89	0.77	375.74	414.81

6 Some more regression caveats

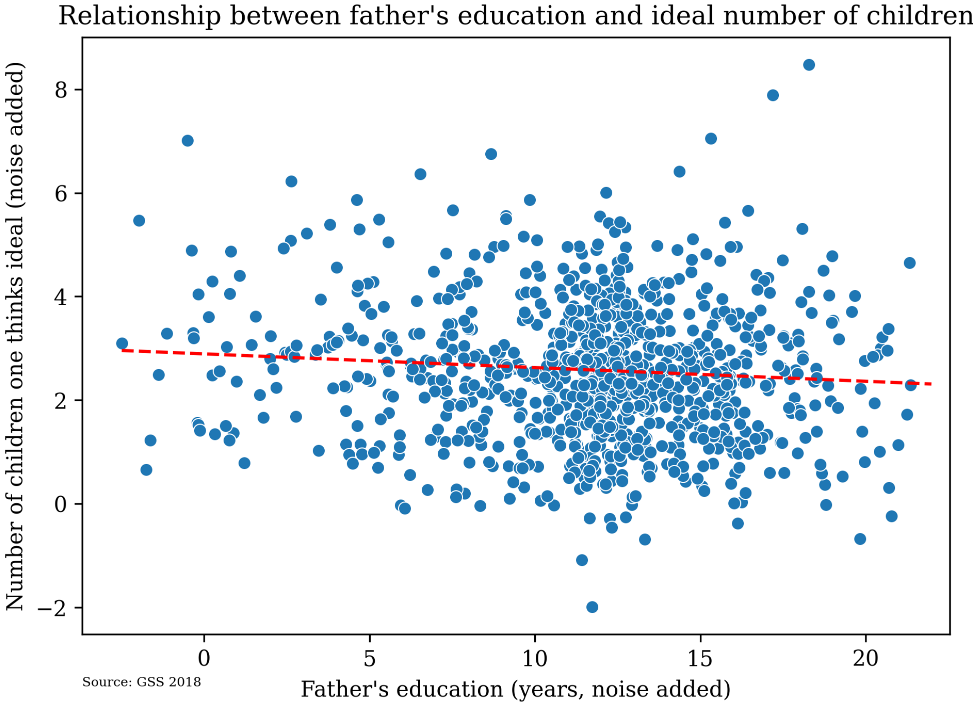
1. **Watch out for outliers**.

   This advice always remains relevant. You should try running analyses with and with outliers. Never simply drop outliers without further effort; this is cheating. Instead, investigate. Our book discusses the use of leverage and influence plots. That gets a little thorny; I'm happy for you to simply investigate points with large residuals. We'll see how to do this later.

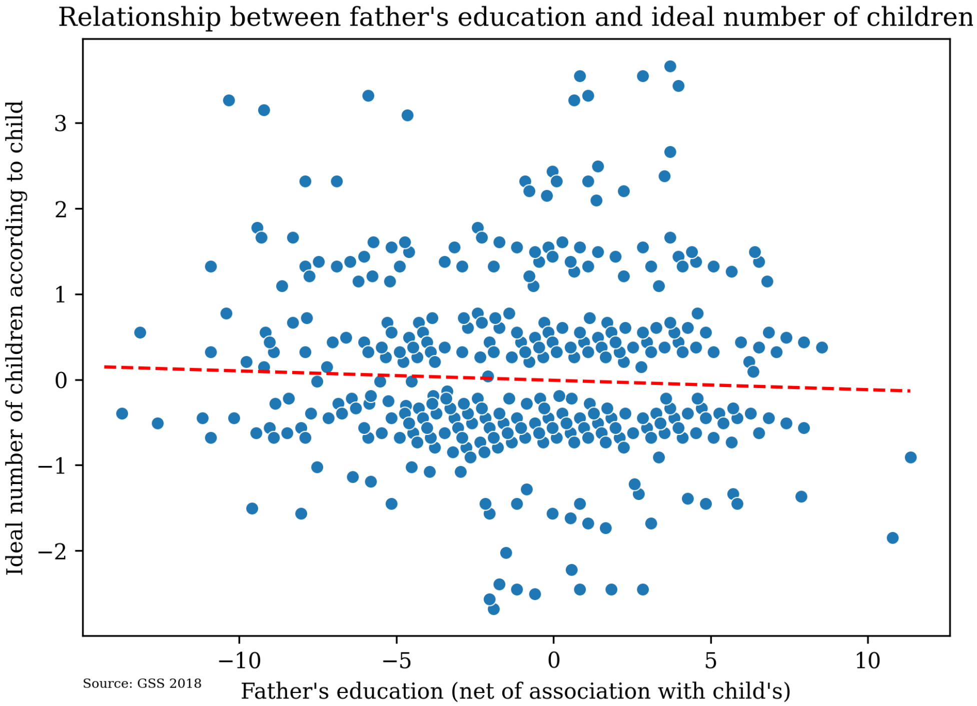
2. **Association is still not causal**.

Remember that our parameter estimates can and will change if we add variables to the model that are relevant and for which we haven't yet accounted.

For example, there is some evidence of a general association between a parent's level of education and how many children someone thinks are ideal. For example, as one's father's education increases, the number of children one thinks are ideal decreases, although this relationship is weak.



However, if we control for the pathway whereby father's education influences a respondent's own level of education, this influence erodes. The conclusion would be that father's education appears to influence the number of children one thinks is ideal by means of influencing how much education a person gets, rather than directly (which makes sense). Notably, although these two graphs might look similar, the slope of the line in the second one is about half as large as that of the former and is *not* statistically significant (more on that later), whereas the previous slope was.



7 Exercises

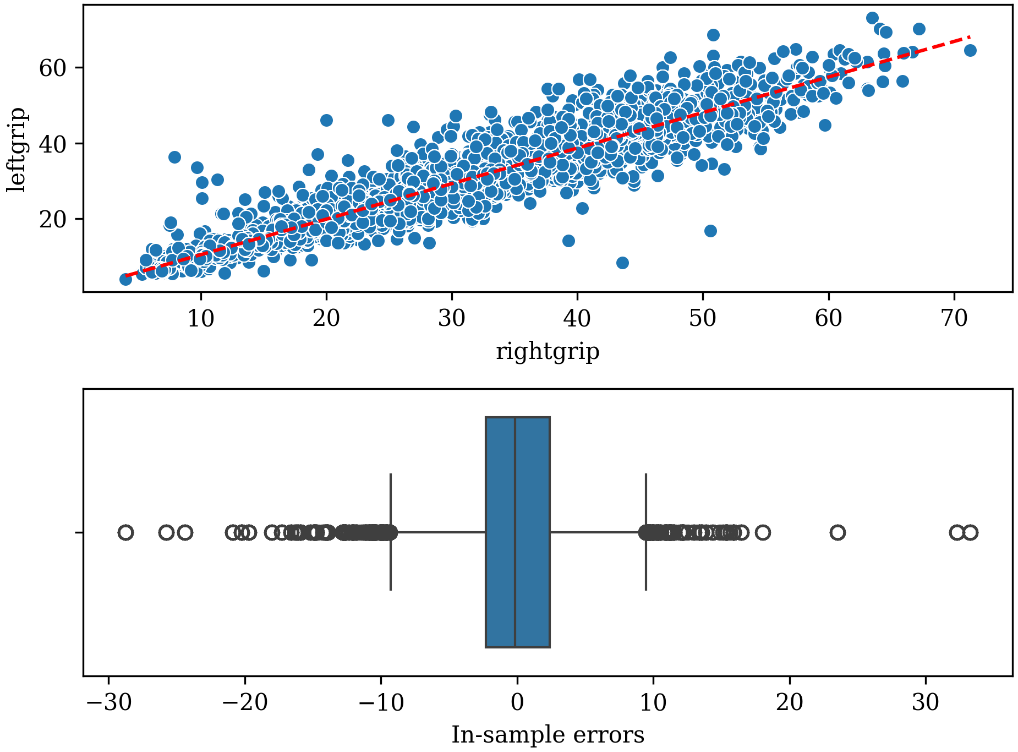
1. Below are NHANES summary statistics for grip strength in respondents' left and right hands (measured in kg) and the correlation matrix for all of the variables (remember: it is symmetric with 1s on the main diagonal of necessity; this is not a typo). Of course, one's left-hand grip strength and right-hand grip strength are closely associated (sociological question: why is the mean a bit higher in the right hand?). Weight and height, on the other hand, are associated with strength, but not as closely. Predict left grip strength from right grip strength, and then predict left grip strength from height. Interpret the equations.

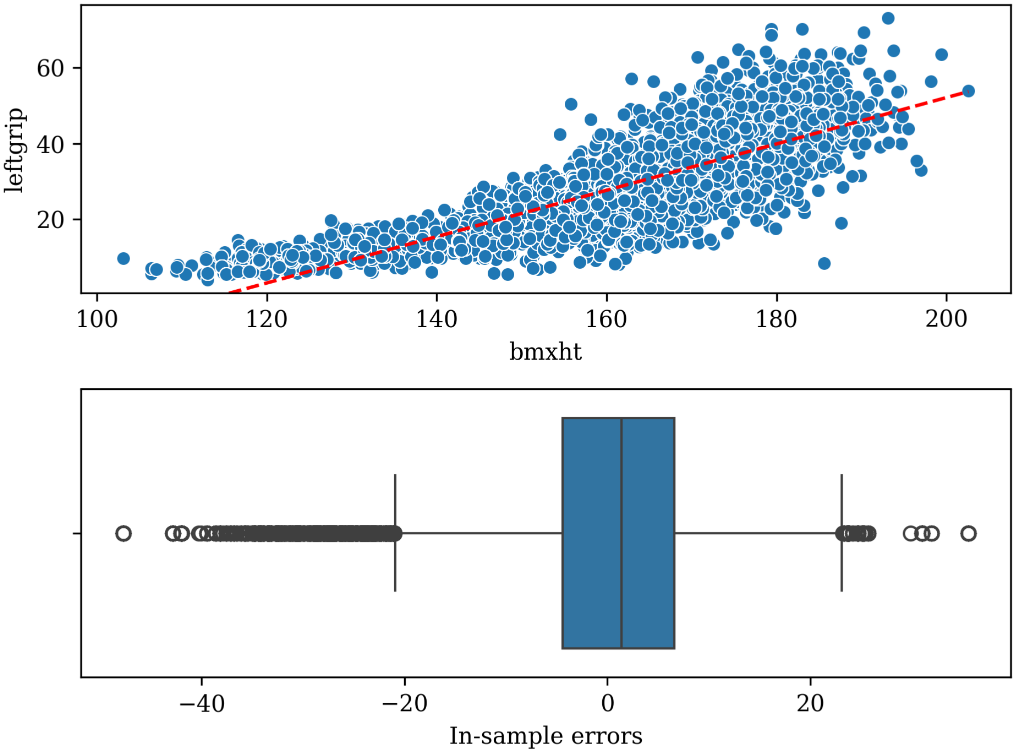
	leftgrip	rightgrip	bmxht	bmxwt
count	53954.00	53954.00	53954.00	53954.00
mean	29.45	30.22	162.95	72.77
std	12.04	11.94	15.56	26.13
min	4.00	4.00	103.10	14.70
25%	21.00	22.00	156.00	56.40
50%	28.10	29.10	164.80	71.90
75%	37.60	38.70	173.30	88.10
max	73.20	71.30	202.60	201.60

	leftgrip	rightgrip	bmxht	bmxwt
leftgrip	1.00	0.93	0.79	0.61
rightgrip	0.93	1.00	0.80	0.60
bmxht	0.79	0.80	1.00	0.70
bmxwt	0.61	0.60	0.70	1.00

2. What share of the variance in left grip strength is explained by variance in right grip strength? How about answering the same question but using height as a predictor?

3. Below is a set of scatterplots characterizing the above situation. Where are the simple linear regression assumptions better met? How can you tell from the residuals?





4. Open Stata. Make a standard do-file (at this point, I'm just going to let you do this on your own). Open the GSS 2018 data. Make a scatterplot of a respondent's mother's and father's education with a small jitter (`scatter maeduc paeduc, jitter(10)`). Regress mother's education on father's education (an usurprising connection) using `reg maeduc paeduc`. Interpret the regression equation. Then, add a regression line like so (I also turn the legend off; it's pretty annoying for scatter plots in Stata).

   ```
   scatter $y$ $x$, jitter(10) || lfit $y$ $x$, legend(off)
   ```

5. Do some diagnostics. First, how plausible is our model mix? We only have one predictor, but was it a good one? How good was our prediction? Look for r^2 and check that $\frac{SS_M}{SS_T} = r^2$. Then, make a histogram and a boxplot of the residuals as follows. Does our linearity assumption look OK, or was the functional form misspecified?

   ```
   reg maeduc paeduc
       * you must do this first; the next command assumes regression results in memory
   predict e_sample, residuals
       * the "residuals" option tells Stata to store them in a column called "e_sample"
   hist e_sample
   graph hbox e_sample
   ```

   Check that the standard deviation of the in-sample errors is the $RMSE$ and that their mean is 0.

6. Do you suspect that any of the outliers have a strong influence on the regression

results?

8 Exercise answers

1. First, find the correlations in the table. For left/right, it is 0.93; for left/height, it is 0.79. Then, multiply by the ratio of standard deviations to get the slope, b_1: $0.93 \cdot \frac{12.04}{11.94} = b_1 = 0.94$ in the first case and $0.79 \cdot \frac{12.04}{15.56} = 0.61$ in the second. Then, find the intercept by finding $b_0 = \bar{y} - b_1\bar{x}$ in each case. For the first case, we have $29.45 - 30.22 \cdot 0.94 = 1.04 = b_0$; for the second, we have $29.45 - 162.95 \cdot 0.61 = -69.95 = b_0$. So, our first model tells us that a kilogram increase in right grip strength is associated with about a 0.94 kg increase in left grip strength. Conversely, a cm increase in height is associated with about a 0.61 kg. increase in strength. Here, the intercepts aren't really meaningful.
2. Here, we can just square the correlation coefficients. The first model explains $0.93^2 = 0.86$, while the second explains $0.79^2 = 0.62$.
3. The model assumptions are fairly plausible in the case of left/right grip strength, but they aren't as plausible in the case of height. The in-sample errors are clearly
4. Our estimated model is $y = 0.63 \cdot x + 4.43$. So, for a given year increase in one's father's education, we expect one's mother's education to increase by 0.63 years. Here, the intercept isn't wildly meaningless: we expect one's mother to have about 4.5 years of education if one's father has 0.
5. Our model mix is pretty good given that we can only use one predictor. The functional form looks alright. Our prediction was pretty good; our model explains 0.47 percent of the variance in the outcome.
6. Here, just eyeballing it, no outlier seems to be driving regression results.

9 Appendix

9.1 Sums of squares formulae

The formulae for the sums of squares above are a little bit tricky. We saw from the geometry that the squared length of the prediction vector for centered data plus the squared length of the error vector for centered data is the squared length of the centered outcome. How do we *generalize* these formulae to the uncentered case? In general, the guiding thread is that we want to make our formulae general so that if we apply them to an uncentered case, we get the quantities from the centered geometric picture—which we *know* work from basic geometry—from our uncentered problem.

First of all, our outcome vector doesn't change. We interpret this as the total sum of squares in both cases; in the centered case, we simply have the option of renaming our variable from something like $y - \bar{y}$ to simply w, but this doesn't really do anything helpful. We interpret the squared length of the outcome as proportional to the variance in both cases.

Note also that the error vector is the same in the uncentered and centered models. In the centered case, we have $err_j = (y_j - \bar{y}) - b_1(x_j - \bar{x})$. Once we've found that *centered* regression equation, we can take the residual for it, which ends up being the same as our centered residual: $err_j = y_j - (b_1 x_j + b_0) = y_j - b_1 x - \bar{y} + b_1 \bar{x}$ clearly rearranges to $(y_j - \bar{y}) - b_1(x_j - \bar{x})$. So, we don't modify that formula either.

Our model sum of squares is a bit different. The predictions in the centered case are $b_1(x_j - \bar{x})$, and we would simply square this quantity and sum to find the SS_M. However, in the centered case, we have $b_1 x_j + b_0 = b_1 x_1 + \bar{y} - b_1 \bar{x} = b_1(x_1 - \bar{x}) + \bar{y}$. We have an extra $\bar{y}$ in the unadjusted predictions, so our SS_M becomes $\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2$.

9.2 Conditional means as the best possible predictor

Let's work out the formula for the best possible function $f(x)$ for the sample data (an analogous result holds in the population). To work this out, instead of summing over individuals directly, let's sum over the possible *values* of the variable, then the number of individuals who take that value. This requires double summation notation, but it's really not so bad, no different than what I just described.

$$\begin{aligned}
\text{MSE} &= \sum_{j=1}^{n_h} \sum_{h=1}^{m} \left\{ y_{hj} - f(x_h) \right\}^2 \\
&= \sum_{j=1}^{n_h} \sum_{h=1}^{m} y_{hj}^2 + \sum_{j=1}^{n_h} \sum_{h=1}^{m} f(x_h)^2 - 2 \sum_{j=1}^{n_h} \sum_{h=1}^{m} y_{hj} f(x_h) \\
&= \qquad \underbrace{\sum_{j=1}^{n_h} \sum_{h=1}^{m} y_{hj}^2}_{\text{doesn't depend on prediction}} \qquad + \sum_{j=1}^{n_h} \sum_{h=1}^{m} f(x_h)^2 - 2 \sum_{j=1}^{n_h} \sum_{h=1}^{m} y_{hj} f(x_h)
\end{aligned}$$

Now, we can focus on trying to minimize the parts that involve the function of our variable. Through just a bit of hand-inspection, we see that this can be made to equal zero if $f(x_h) = \bar{y}_j$, i.e. the conditional mean of y at $x = x_h $.

$$\text{MSE} = \sum_{h=1}^{m} n_h f(x_h)^2 - 2 \sum_{h=1}^{m} n_h \bar{y}_h f(x_h)$$

9.3 Correlation coefficient and regression slope connection

$$\begin{aligned}
r \cdot \frac{s_y}{s_x} &= r \cdot \frac{\sqrt{\frac{\sum_{j=1}^{n}(y_j-\bar{y})^2}{n-1}}}{\sqrt{\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1}}} \\
&= \frac{\sum_{j=1}^{n}(y_j-\bar{y})(x_j-\bar{x})}{\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_j-\bar{y})^2}} \frac{\sqrt{\sum_{j=1}^{n}(y_j-\bar{y})^2}}{\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \\
&= \frac{\sum_{j=1}^{n}(y_j-\bar{y})(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} \\
&= b_1
\end{aligned}$$